Tools for Assessing Situational Awareness in an Operational Fighter Environment

WAYNE L. WAAG, B.S., Ph.D., and MICHAEL R. HOUCK, M.S., Ph.D.

WAAG WL, HOUCK MR. *Tools for assessing situational awareness in an operational fighter environment.* Aviat. Space Environ. Med. 1994; 65(5, Suppl.):A13–9.

Three Situational Awareness Rating Scales (SARS) were developed to measure pilot performance in an operational fighter environment. These instruments rated situational awareness (SA) from three perspectives: supervisors, peers, and self-report. SARS data were gathered from 205 mission-ready USAF F-15C pilots from 8 operational squadrons. Reliabilities of the SARS were quite high, as measured by their internal consistency (0.95 to 0.99) and inter-rater agreement (0.88 to 0.97). Correlations between the supervisory and peer SARS were strongly positive (0.89 to 0.92), while correlations with the self-report SARS were positive, but smaller (0.45 to 0.57). A composite SA score was developed from the supervisory and peer SARS using a principal components analysis. The resulting score was found to be highly related to previous flight experience and current flight qualification. A prediction equation derived from available background and experience factors accounted for 73% of its variance. Implications for use of the composite SA score as a criterion measure are discussed.

SITUATIONAL AWARENESS (SA) has generated considerable interest within the aviation community. Loss of SA has been considered a major contributory factor in many military and commercial aviation accidents and incidents (18). The human factors community is pursuing the measurement of SA as a tool for evaluating cockpit design (14). There is also interest in SA within the fighter pilot community, since it is considered a key element in determining success during air combat operations (21).

In 1991, the U.S. Air Force Chief of Staff posed a series of questions concerning SA that led to the present investigation. First of all, what is SA? Can it be objectively measured? Is SA learned or does it represent a basic ability or characteristic that some pilots have and others do not? From a research standpoint, these questions translate into issues of measurement, selection, and training. Armstrong Laboratory was subsequently tasked with providing research answers to these questions. A research investigation was initiated that had three goals: first, to develop and validate tools for reliably measuring SA; second, to identify basic cognitive and psychomotor abilities that are associated with pilots judged to have good SA; and third, to determine if SA can be learned, and if so, to identify areas where cost-effective training tools might be developed and employed.

The general approach was first to develop criterion measures of SA based upon performance ratings collected within an operational flying environment. These measures were necessary for two reasons. First, they would serve as criterion measures against which to validate a battery of basic ability tests considered relevant to SA, thereby addressing the question of basic human abilities. Second, these measures would serve as a means of selecting a sample of pilots who would participate in a simulation phase of the effort. During that phase, simulated air combat mission scenarios were developed for assessing SA, and objective measures of performance gathered in an attempt to determine those characteristics that distinguish pilots with good SA. These data would be used to identify areas where training tools might be developed. This article presents the results of only the first phase of the program; namely, the development of tools for measuring SA within an operational fighter environment.

The approach to developing measurement tools was largely dictated by the definition of SA adopted at the outset of the study, the intended use of the data, and practical constraints involved in gathering data on mission-ready aircrew. In response to the question, "What is SA?", the Air Staff produced the following operator's definition: "a pilot's continuous perception of self and aircraft in relation to the dynamic environment of flight, threats, and mission, and the ability to forecast, then execute tasks based on that perception" (2). While other definitions of SA within the literature focus pri-

From Armstrong Laboratory, Human Resources Directorate, Mesa, AZ (W. L. Waag) and University of Dayton Research Institute, Dayton, OH (M. R. Houck).
Address reprint requests to: Dr. Wayne L. Waag, AL/HRAT, 6001 S. Power Rd., Bldg. 558, Mesa, AZ 85206-0904.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
1994

2. REPORT TYPE

3. DATES COVERED

4. TITLE AND SUBTITLE
Tools for Assessing Situational Awareness in an Operational Fighter Environment

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Wayne Waag; Michael Houck

5d. PROJECT NUMBER
1123

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Armstrong Laboratory,Operations Training Division,6001 South Power Road,Mesa,AZ,85206-0904

8. PERFORMING ORGANIZATION REPORT NUMBER
AL; AL/HRA

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Three Situational Awareness Rating Scales (SARS) were developed to measure pilot performance in an operational fighter environment. These instruments rated situational awareness (SA) from three perspectives: supervisors, peers, and self-report. SARS data were gathered from 205 mission-ready USAF F-15C pilots from eight operational squadrons. Reliabilities of the SARS were quite high, as measured by their internal consistence (0.95 to 0.99) and inter-rater agreement (0.88 to 0.97). Correlations between the supervisory and peer SARS were strongly positive (0.89 to 0.92), while correlations with the self-report SARS were positive, but smaller (0.45 to 0.57). a composite SA score was developed from the supervisory and peer SARS using a principal components analysis. The resulting score was found to be highly related to previous flight experience and current flight qualification. A prediction equation derived from available background and experience factors accounted for 73% of its variance. Implications for use of the composite SA score as a criterion measure are discussed.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**7**	

marily on processes underlying the assessment of the situation (13), our working definition also included forecasting, decision-making, and task execution. As such, it was viewed as a fairly global operational concept that encompasses much of the domain of air combat proficiency.

Since the data were to be used primarily as a criterion against which to determine relationships with basic ability measures, fairly large numbers of subjects would be required. This requirement further restricted the types of measures to those that could be gathered in a quick, noninvasive manner, since available pilot time within any operational flying environment is limited. While a number of measurement tools had been developed to measure SA within a highly controlled, simulated flight environment (4,5), these did not seem appropriate for the present study. Rather, previous efforts to develop criterion measures of combat effectiveness seemed more germane (22).

Attempts to measure and predict combat effectiveness have a long history dating back to World War II. The interested reader is referred to Youngling et al. (22), who conducted an extensive review of this literature. In essence, there are two problems that must be addressed: first, the definition of the criterion for combat effectiveness; and second, the search for measures that are predictive of that criterion. In general, four types of criteria have been used: objective outcome measures such as kills, bombing scores, etc.; direct and systematic observations of mission performance; administrative actions such as failure to complete a fighter tour; and qualitative ratings of overall ability. On the predictor side, a variety of potential indicators of combat effectiveness have been explored, including basic aptitude, biographical factors including flight experience, a variety of personality and motivational factors, perceptual-motor abilities, and knowledge and skills directly related to aviation.

In general, only very modest relationships have been obtained. Of the predictor sets that have been evaluated, those measures related to previous flight experience seem to be most consistently related to combat effectiveness as measured by combat kills. Strawbridge and Kahn (17) and Torrance et al. (20) reported correlations with previous flight experience in the range of 0.30 to 0.40. Correlations with aptitude test scores and other perceptual-motor tests were substantially lower, with most failing to reach statistical significance. DeLeon (3) summarized the results of the Red Baron studies that were conducted during the Vietnam conflict. Flight experience, in terms of total flight hours, total fighter hours, and hours in the combat aircraft, was found to be related to combat success, although the degree of these relationships was fairly small. DeLeon (3, p. 16) concluded that "at best, experience appears to be only a vague measure of a pilot's air-to-air combat skills."

In summary, previous studies have reported the highest relationships between flight experience and criteria involving actual combat success; i.e., kills. However, such criteria were not available for the present study; nor was it feasible for reasons of time, cost, and lack of combat realism to gather data based on actual mission performance in the aircraft under the highly controlled conditions of an instrumented range environment, as suggested by Youngling et al. (22). For practical reasons, the only alternative was to develop criterion measures based upon human judgments. Unfortunately, the use of subjective ratings of overall ability as the criterion of combat effectiveness has produced few statistically significant relationships with predictor sets that have been used to date. For example, Lepley (12) found only one significant correlation for his test battery with subjective ratings of ability. Shannon and Waag (15,16) met with limited success in an attempt to relate background and experience factors to operational performance. In this case, squadron commander ratings of mission-critical performance dimensions were the criterion measures. Results indicated that flight experience was the best predictor of criterion performance. Undergraduate Pilot Training (UPT) grades for formation and tactics were also found related to such ratings. However, the overall magnitude of the relationship was fairly low with a multiple correlation of all background and experience factors reaching only 0.35.

In general, three types of performance ratings have been used. The most common, and also most cited in the literature, has been supervisory ratings (11). The two other types include peer ratings and self ratings. In fact, the use of peer ratings for combat aviation dates back to World War II, when Jenkins et al. (7) developed a "combat" criterion for the U.S. Navy, based largely upon peer nominations. In an extensive literature review, Landy and Farr (11) conclude that previous research studies have not found very high correlations among these three types of ratings. Moreover, it is difficult to select one approach as best since the literature does not suggest any of these to be more valid than the others. For these reasons, it was decided to develop three SA Rating Scales (SARS) and gather supervisory, peer, and self-report data. Moreover, it was decided to use a simple graphic scale since the literature is equivocal regarding more elaborate procedures such as behaviorally anchored rating scales (11).

What seemed most critical, however, were the actual dimensions that were to be rated and the development of clear definitions for each. To characterize the domain of air combat, it was necessary first to identify and describe the critical activities required of the fighter pilot to maintain good SA and complete his mission successfully. To this end, Houck et al. (10) conducted a cognitive task analysis of the attack portion of an F-15C air combat mission. This analysis relied primarily on the input of experienced fighter pilots and focused on critical air combat task categories that in previous research were rated by F-15C pilots as being most amenable to training in air combat simulations (8,9,19). The resulting analysis identified the significant types of decisions required of the flight members, the information required for making these decisions, and the observable activities the flight members performed to acquire this information. For the purposes of the present research, this classification provided a detailed description of optimum performance in air combat.

The resulting classification was further analyzed by an experienced fighter pilot to derive those aspects of

air combat operations judged most essential to SA. Paramount in this selection process was that the items must be observable in the context of day-to-day squadron training activities and subject to evaluation by other fighter pilots both in terms of their own performance and that of others. A further requirement was that the pilots must be able to assess these items in retrospect, based on performance observed to date. As a result of this analysis, 24 items organized in 7 categories were produced. Categories included tactical game plan, system operation, communication, information interpretation, beyond-visual-range weapons employment, visual maneuvering, and general tactical employment. Because the 24 items were heavily weighted toward specific operational tasks, an additional 7 items were included to reflect more general traits which also were hypothesized to play a role in SA. These items were based on the study of fighter pilot combat effectiveness previously discussed (22). Concise definitions for each item were developed with assistance from an experienced fighter pilot. The resulting list and definitions were reviewed and revised by several other experienced pilots to ensure accuracy and completeness. These 31 items and the 8 categories that they represent are presented in Table I and form the essence of the approach taken to the measurement of SA in the present study.

To summarize, the purpose of the present investigation was to develop a set of tools for measuring SA within an operational fighter environment. Issues to be addressed include: 1) reliability of the SARS and potential effects of bias factors such as flight qualification of the rater and squadron membership; 2) inter-relationships among the supervisory, peer, and self-report SARS; 3) development of a single composite SA score; and 4) external validity of the composite SA score as determined by relationships with previous flight experience factors.

METHODS

Subjects

The subjects were 205 mission-ready USAF F-15C pilots from 8 operational fighter squadrons. Mean, standard deviation, and range of flight hours, respectively, were as follows: total flight hours beyond UPT (1258, 744, 202 to 3717) and total flight hours in the F-15 (668, 305, 74 to 1823). Current flight qualifications of these pilots, in order of increasing experience and proficiency, included 48 wingmen, 59 2-ship leads, 32 4-ship leads, and 66 instructor pilots.

Materials

Three scales were developed to measure the SA ability of a pilot from three different perspectives: the self-report SARS, the peer SARS, and the supervisory SARS. Survey forms were custom-designed and reproduced through an offset printing process to make use of computer-based data scanning technology. Each survey type was two pages: the first page contained printed instructions, scale description, and subject identification codes; the second page, the actual rating scales.

For the self-report SARS, subjects rated their own ability on each of the 31 items in comparison with other F-15C pilots using a 6-point scale. End-point anchors ranged from a low of "Acceptable," since all pilots were mission-ready, to a high of "Outstanding." The peer SARS required each subject to rate all other mission-ready pilots in his squadron. Each pilot listed on the peer SARS was rated on his general fighter pilot ability and SA ability using the same 6-point scale. Once these ratings were completed, these pilots were then rank-ordered from highest to lowest in terms of their SA ability. A provision was included on the form for not rating a pilot if the rater felt he had insufficient knowledge of that particular individual. The supervisory SARS used the same 31 items and the 6-point scale as the self-report SARS. Again, the reference was the relative ability of the ratee in comparison with other F-15C pilots.

The self-report and peer SARS were completed by all subjects within the sample. The supervisory SARS were completed by only a subset of subjects chosen to be raters, based upon their experience and supervisory positions. Raters within each squadron included: the Squadron Commander, Ops (Operations) Officer, As-

TABLE I. ITEMS AND CATEGORIES USED IN SARS.

1. GENERAL TRAITS	5. INFORMATION INTERPRETATION
Discipline	Interpreting vertical situation display
Decisiveness	Interpreting threat warning system
Tactical knowledge	Ability to use controller information
Time-sharing ability	Integrating overall information
Spatial ability	Radar sorting
Reasoning ability	Analyzing engagement geometry
Flight management	Threat prioritization
2. TACTICAL GAME PLAN	6. TACTICAL EMPLOYMENT-BVR
Developing plan	Targeting decisions
Executing plan	Fire-point selection
Adjusting plan on-the-fly	7. TACTICAL EMPLOYMENT-VISUAL
3. SYSTEM OPERATION	Maintain track of bogeys/friendlies
Radar	Threat evaluation
Tactical electronic warfare system	Weapons employment
Overall weapons system proficiency	8. TACTICAL EMPLOYMENT-GENERAL
4. COMMUNICATION	Assessing offensiveness/defensiveness
Quality (brevity, accuracy, timeliness)	Lookout
Ability to effectively use information	Defensive reaction
	Mutual support

sistant Ops Officer, Weapons Officer, and Stan-Eval (Standard-Evaluation) Flight Examiner (SEFE) who rated all mission-ready pilots within the Squadron; and the Flight Commanders, who rated only pilots within their flight as well as other Flight Commanders.

Procedures

The surveys were administered on location at each fighter squadron base. An elaborate numerical coding procedure was followed to ensure the confidentiality of each subject's data. The survey administrators briefed all subjects regarding the objectives of the research, scale description and item definitions, confidentiality procedures, and instructions for completing the surveys. Identification codes and dates on each survey were already filled in for each subject prior to administration. Each subject removed the surveys enclosed within the envelope, completed them, returned them to the envelope, and removed all name labels. These labels were given to the test administrator who destroyed them, thus leaving no name identification within or outside the envelope.

Data regarding each subject's flight career experience were obtained directly from a computerized database and through responses to a background questionnaire administered to each subject. These data included flight hours and sorties by aircraft type, hours and sorties for both combat and combat support missions, current flight qualification, supervisory responsibilities, advanced fighter training, and participation in special fighter exercises and training simulations.

For the self-report SARS, nine summary scores were produced. These included an overall score, which was the mean of all 31 items, and 8 category scores, which were the means of all items within a particular category. For the supervisory SARS, the same nine summary scores were generated for each ratee as follows: first, the same nine summary scores were computed for each rater's assessment of each ratee; then, means for each summary score were computed across all raters and used as the final nine supervisory SA scores for each ratee. For the peer SARS, three summary scores were produced for each ratee as follows: first, three scores were generated by each rater for each ratee, the ratings of fighter pilot ability and SA ability, and the rank order; means of these three scores were then computed across all raters and used as the final peer SA scores for each ratee.

RESULTS

SARS reliability: The first set of analyses addressed the reliability of the three SARS instruments and checked for systematic biases due to the current flight qualification of the rater or squadron membership. Two types of reliability were estimated: internal consistency and inter-rater agreement. First, internal consistency was estimated for the supervisory and self-report SARS by calculation of Cronbach's coefficient α. For the supervisory SARS, coefficient α was computed to be 0.99 for all 31 items. These results were based on the total number of supervisory SARS completed (N = 884). For the self-report SARS, α was computed to be 0.97 for all 31 items. Again, these were based on the total number of self-report SARS completed (N = 187). Second, inter-rater agreement was estimated for the supervisory and peer SARS. Scores used in the calculation of these estimates included the nine scores from the supervisory SARS (eight category scores plus overall score) and the three scores from the peer SARS. Reliability estimates for each of the 12 scores were computed for each squadron, using an analysis of variance (ANOVA) procedure (6). Two estimates of reliability were produced, first the estimated reliability for a single rater, $\bar{r}_{11}$, and second, the reliability of all raters, r_{kk}. For the supervisory SARS, the average $\bar{r}_{11}$, across the eight squadrons was computed to be 0.50. The average r_{kk} increased to 0.88. For the peer SARS, average $\bar{r}_{11}$, was computed to be 0.60 and average r_{kk} increased to 0.97. These data clearly demonstrate the increase in the reliability of the scores through the addition of multiple raters.

A 2-factor ANOVA was used to determine any bias effects in the supervisory and peer SARS due to rater qualification and squadron membership. Eleven ANOVA's were computed for the nine supervisory SARS scores and two of the peer SARS scores, ratings of fighter pilot and SA ability. For each effect, ω^2 was also computed as an estimate of the strength of the association. For rater qualification, no significant effects were found. For squadron membership, only the SA ability rating produced a significant effect with $F(7,151) = 2.08$, $p < 0.05$. ω^2 was computed to be 0.039, indicating a very small effect size. Moreover, of the 28 pairwise comparison tests, only 2 reached significance. Additionally, no significant interaction effects were found for any of the 11 scores.

Similar analyses were conducted for the nine summary scores from the self-report SARS. In this case, however, it was expected that significant effects would occur, due to rater qualification. In other words, it was hypothesized that the self-ratings of instructor pilots, for example, would be higher than the self-ratings of inexperienced wingmen. The results of the ANOVA's confirmed these expectations. All F-ratios were statistically significant ($p < 0.05$). For the overall score, ω^2 was computed to be 0.16, indicating a moderate amount of the variance explained by this factor. Means for all scores were in accordance with expectations and ranked according to qualification with IP's having the highest scores and followed in order by 4-ship leads, 2-ship leads, and wingmen. Significant squadron effects were also obtained for six of the scores including the overall score. Pairwise comparison tests revealed that one squadron accounted for these differences with higher means for all scores. When ANOVA's were recomputed excluding data from that squadron, only one of the nine scores produced a significant squadron effect, accounting for less than 4% of the variance. Only one of the pairwise comparison tests reached significance.

SARS intercorrelations: The second set of analyses computed intercorrelations among the three sets of SARS scores, which are presented in Table II. For the sake of brevity, only correlations with the overall score are presented for both the self-report and supervisory SARS. The average correlation of category scores with the overall score was computed to be 0.95 for the su-

TABLE II. CORRELATIONS AMONG SUPERVISORY, PEER, AND SELF-REPORT SCORES.

	1	2	3	4	5
1. Supervisory SARS-Overall Score	—				
2. Peer SARS-Fighter Pilot Ability	0.89	—			
3. Peer SARS-SA Ability	0.91	0.98	—		
4. Peer SARS-Rank Order	0.92	0.91	0.92	—	
5. Self-Report SARS-Overall Score	0.45	0.56	0.57	0.49	—

pervisory SARS and 0.86 for the self-report SARS, indicating a high degree of internal consistency. All correlations were statistically significant ($p < 0.01$) and, as seen in Table II, the relationships among the supervisory and peer ratings were quite high. Although the correlations of the self-report SARS with the other ratings were positive, their magnitude was substantially lower.

SARS composite score development: In developing a composite SARS score, it was decided to exclude the self-report SARS for two reasons. First, the self-report SARS was significantly influenced by squadron membership. And second, only moderate correlations were found with the supervisory and peer ratings. Consequently, only the three peer SARS scores and the eight supervisory SARS category scores were included in the development of a single composite score. The overall score from the supervisory SARS was excluded since, mathematically, it represented a linear combination of the category scores. A principal components analysis was performed to determine the underlying structure of these scores. The first principal component was found to account for 92.5% of the total variance of these scores, the second component 3.3%, and the remaining components less than 1% of the variance. Based upon these results, it was decided to compute composite scores based upon the first unrotated principal component due to the large amount of variance it explained. These scores were transformed to a distribution with mean of 100 and a standard deviation of 20 for use as the composite SA score in subsequent analyses.

Effects of previous experience on composite SARS score: Analyses were performed to determine if the composite SA score was related to previous flight experience information. It seemed reasonable to expect that measures of experience such as flight hours, flight qualification, and combat training exercise participation should be related, to some extent, to our composite score. In fact, if such relationships were not found, it would seriously question the validity of our composite SA score. Experience factors that were analyzed included: total flight time; total flight time in the F-15; exercise participation (i.e., number attended) including Red Flag (0, 1, 2, ≥3), Green Flag (0, ≥1) Maple Flag (0, ≥1) and William Tell (0, ≥1); air combat simulation training experience (yes/no) including the McDonnell-Douglas Advanced Air Combat Simulation (MACAIR) and the Simulator for Air-to-Air Combat (SAAC); Desert Storm experience (yes/no); and current flight qualification, including whether the pilot was a Fighter Weapons School graduate (yes/no). Additionally, the effect of squadron membership was also analyzed. A one-way ANOVA was computed for each factor, except for flight time. For total flight hours and flight hours in the F-15, correlations were computed. The results are summarized in Table III.

As shown in Table III, most of the experience factors were related to the composite measure of SA. In fact, only two of the measures were not significantly related to SA, participation during Desert Storm and previous training in the Simulator for Air-to-Air Combat. It should also be noted that squadron membership had no effect on the composite SA score. In all cases, the direction of the means was such that higher experience was associated with better SA scores. In fact, some of the relationships were extremely high. For example, current flight qualification accounted for 68% of the variance of the SA measure. These means are presented in Fig. 1. As shown, there is a very strong relationship with flight qualification.

In the final set of analyses, a prediction equation was derived for the composite SA score using a combination of background and experience factors. A stepwise regression analysis was performed with the composite SA score as the dependent variable and those statistically significant background experience factors listed in Table III as the potential set of predictor variables. A "dummy variable" coding scheme was employed to enable entry of flight qualification which is a categorical variable. A 4-variable, "best fit" prediction equation was produced with a multiple R of 0.85. Variables included in the equation, in their order, were flight qualification, graduation from Fighter Weapons School, participation at Green Flag, and participation at Maple Flag. The overall multiple R was statistically significant ($p < 0.0001$) as well as the contribution of each variable within the equation ($p < 0.05$).

DISCUSSION

Three measurement tools were developed for assessing SA within an operational fighter environment. The primary concerns with any measurement device are its reliability, susceptibility to unwanted bias factors, and its validity.

SARS reliability: Reliability estimates, in all cases, were quite high. Estimates of internal consistency for both the self-report and supervisory SARS exceeded

TABLE III. EFFECTS OF BACKGROUND AND EXPERIENCE FACTORS ON COMPOSITE SA SCORE.

	F-Ratio	p	$\bar{\omega}^2$
Squadron	0.59	NS	0.00
Flight Qualification	128.57	<0.001	0.68
Exercise Participation			
Red Flag	13.55	<0.001	0.18
Green Flag	6.15	<0.01	0.06
Maple Flag	5.28	<0.05	0.02
William Tell	17.05	<0.001	0.09
Fighter Weapons Grad	55.85	<0.001	0.24
Simulator Experience			
MACAIR	29.81	<0.001	0.14
SAAC	1.01	NS	0.00
Desert Storm Veteran	1.54	NS	0.00
F-15 Hours	0.59*	<0.001	0.35
Total Flight Hours	0.39*	<0.001	0.15

* Correlations.

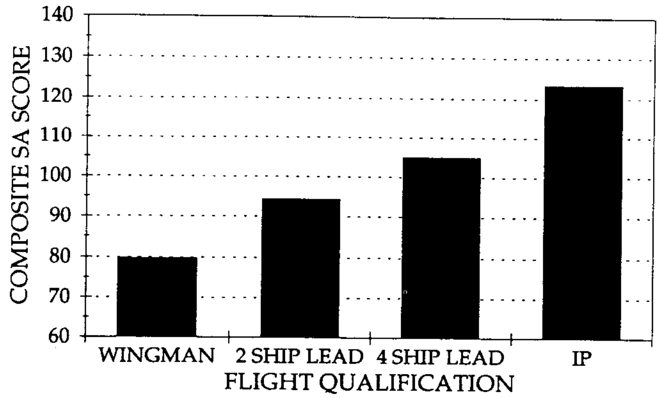

Fig. 1. Composite SA Score as a Function of Flight Qualification.

0.95, indicating that whatever the 31 items might be measuring, they are indeed measuring it consistently. Of greater importance, however, are the estimates of inter-rater reliability. It was reasoned that both the reliability and validity of the criterion SA scores would be enhanced if each ratee were evaluated by multiple raters. Consequently, for the supervisory SARS, each ratee score was based on an average of from five to eight raters. For the peer SARS, the numbers ranged from 18 to 23. The results of the reliability analyses confirm the value of such an approach. The average reliabilities across squadrons obtained for a single rater for both the supervisory and peer SARS were marginal. However, a large increase occurred when the average scores for all raters were used as the estimate. Although such increases in reliability from use of multiple raters seem intuitive, the performance rating literature (11) has not always produced such effects.

Sources of bias: In addition to reliability, there was also concern that the SARS might be systematically biased. The two major potential sources of such bias were squadron membership and the qualification of the rater. The results indicated that no such significant effect occurred for either the supervisory or peer SARS. Of the 33 effects tested, only 1, the peer rating for SA ability, produced a statistically reliable difference. However, not much significance is attached to that difference for two reasons. First, the size of the effect was quite small, accounting for less than 4% of the variance. Second, the peer rating for fighter ability produced no significant difference despite the fact that it was highly correlated ($r = 0.98$) with the peer rating of SA ability. For the self-report SARS, however, significant effects due to rater qualification were expected, and these were confirmed by the results. Of interest was the finding that one squadron produced significantly higher ratings. When the data from this squadron were removed, no meaningful squadron effect was obtained. Reasons for the elevated self-report SARS scores for the one squadron were not apparent.

Interrelationships among SARS: An analysis of interrelationships among the SARS scores produced extremely high correlations between the supervisory and peer SARS scores. Such magnitude would not have been expected, based on the previous literature (11). Of greater consistency with the literature were the relationships with the self-report SARS scores. Although positive correlations were obtained between the self-report SARS and the supervisory and peer SARS, their magnitudes were significantly lower. Moreover, a comparison of the overall means between the supervisory and self-report SARS revealed higher means for the self-report SARS scores, which is consistent with the previous findings of a "leniency" effect of self-ratings (11).

The high degree of consistency between the supervisory and peer SARS scores was further confirmed by the principal components analysis in which the first component accounted for over 92.5% of the total variance. The average correlation between the eight category SARS scores and the first component score was 0.96. The second component accounted for an additional 3.3% of the variance and seemed to represent some unique variance associated with the peer SARS. Correlations with the component score were 0.34 and 0.33 for fighter pilot and SA ability, respectively, and 0.19 for the ranking. All correlations with the supervisory SARS scores were negative and most (six of eight) were not statistically significant. Overall, these results further substantiate the high agreement between the supervisory and peer SARS score and the existence of a very large component that can account for most of the variance. Although there appears to be a second component that is capturing some unique variance associated with the peer SARS, its size was quite small, and consequently not used as a criterion measure of SA.

Effects of flight experience: At the outset, we hypothesized that there would be positive relationships between flight experience and the SA criterion measure. In fact, any measure that was unrelated or negatively related to flight experience would be highly suspect. The results clearly supported our hypotheses in that most of the experience data produced positive relationships with the composite SA score. The finding of positive correlations of both total flight time and time in the F-15 is consistent with the earlier literature, although the obtained correlations were higher than had been reported previously. Current flight qualification was the variable found most highly correlated with the criterion SA measure. In fact, this variable alone accounted for nearly 68% of the criterion variance. When flight qualification was combined with other available information, a prediction equation could be developed which accounted for nearly 73% of the criterion variance, which is equivalent to a correlation of 0.85. These results clearly indicate that the criterion measure of SA developed for this study can be predicted reasonably well from readily available background and flight experience information.

Interpretation and use: Two questions emerge from these findings. First, what is actually being measured by the SARS? And second, what are implications for use of the composite SARS score as a criterion measure? An inherent problem with most criterion measures is that they usually represent a "picture" in time (1). Within the operational fighter environment, pilots progress in a fairly "lock step" manner as they move from one flight qualification to another. F-15 pilots begin their careers in an operational fighter squadron by completing mission qualification training. At that point they are des-

ignated mission-ready wingmen. After a certain number of hours in the jet, they become eligible for upgrade to 2-ship lead. If successful, they gain experience (i.e., flight hours) and eventually become eligible for upgrade to 4-ship lead. Within this process, a certain amount of "selection" occurs. If they are judged not to have the requisite skills for upgrade, their careers as fighter pilots will usually end and they will be reassigned. Viewed in this manner, it is not surprising that current flight qualification is highly related to our criterion measure of SA. It is clear that all raters (both supervisors and peers) were aware of each ratee's flight qualification within the squadron. Such knowledge likely provided a good frame of reference and to some unknown extent may have been the basis for making judgments required in the SARS. Consequently, it appears that the SARS, in large part, measures what might be termed an Air Force "operational management" view of fighter pilot skill, and as such would be highly correlated with flight experience and current qualification. However, the criterion SA measure is more than "experience only," as indicated by the number of what might be termed exceptions. For example, instances occurred in which an individual's criterion SA score was "inconsistent" with his or her qualification level, such as IP's who received scores more characteristic of wingmen. And conversely, there were some wingmen and 2-ship leads that received scores much higher than their experience would suggest.

Implications for use of the composite SA score as a criterion measure are fairly straightforward. It is clear that effects due to background and flight experience must be controlled when these scores are used as criterion measures. This could be accomplished by partialling out these effects and using scores representing the residual variance as measures. Alternatively, separate analyses could be conducted for each qualification category. Regardless, the fact remains that experience accounted for a very large percentage of the variance within our criterion measure. It could be argued that, within the highly homogenous and select population of the F-15 fighter community, training and experience are likely to account for more variability than ability differences. It is expected that results of the simulation phase of this study will shed additional light on this issue.

ACKNOWLEDGMENTS

The authors wish to thank a number of individuals who were instrumental in the accomplishment of this study. In particular, Mr. Rocky Kendall, a retired USAF fighter pilot currently employed by the Greystone Corporation, provided much of the required subject matter expertise in the identification of items relevant to SA and the development of standard definitions. Mr. Bart Raspotnik and Mr. David Greschke, also retired USAF fighter pilots and currently with the University of Dayton Research Institute (UDRI), served as test administrators. Dr. David Hubbard and Ms. Elzbieta Jackiewicz, also of UDRI, provided statistical assistance. And finally, the authors wish to express their appreciation to the 8 fighter squadrons and their personnel who provided the "data" on which these results are based. Without their dedication and support, the accomplishment of this study would have been impossible.

REFERENCES

1. Austin JT, Villanova P. The criterion problem: 1917–1992. J. Appl. Psychol. 1992; 77:836–74.
2. Carroll LA. Desperately seeking SA. TAC Attack (TAC SP 127-1) 1992; 32:5–6.
3. DeLeon P. The peacetime evaluation of the pilot skill factor in air-to-air combat. Santa Monica, CA: The Rand Corporation, 1977; R-2070-PR.
4. Endsley MR. Predictive utility of an objective measure of situational awareness. Proceedings of 34th Annual Meeting. Santa Monica, CA: Human Factors Society, 1990:41–5.
5. Fracker ML. Measures of situation awareness: review and future directions. Wright-Patterson AFB, OH: Armstrong Laboratory, 1991; AL-TR-1991-0128.
6. Guilford JP. Psychometric methods. New York: McGraw-Hill, 1954.
7. Jenkins JG, Ewart ES, Carroll JB. The combat criterion in naval aviation. Washington, DC: National Research Council on Aviation Psychology, 1950.
8. Houck MR, Thomas GS, Bell HH. Training potential of multiplayer air combat simulation. Proceedings of 33rd Annual Meeting. Santa Monica, CA: Human Factors Society, 1989: 1300–4.
9. Houck MR, Thomas GS, Bell HH. Training evaluation of the F-15 advanced air combat simulation. Brooks AFB, TX: Armstrong Laboratory, 1991; AL-TP-1991-0047.
10. Houck MR, Whittaker LA, Kendall RR. An information processing classification of beyond-visual-range intercepts. Brooks AFB, TX: Armstrong Laboratory, 1993; AL/HR-TR-1993-0061.
11. Landy FJ, Farr JL. Performance rating. Psychol. Bull. 1980; 87: 72–107.
12. Lepley WM, ed. Psychological research in the theaters of war. Washington, DC: U.S. Government Printing Office, 1947; AAF Psychol. Prog. Res. Report No. 17.
13. Sarter NB, Woods DD. Situational awareness: a critical but ill-defined phenomenon. Int. J. Aviat. Psychol. 1991; 1:45–57.
14. Selcon SJ, Taylor RM, Koritsas E. Workload or situational awareness? TLX vs. SART for aerospace system design evaluation. Proceedings of 35th Annual Meeting. Santa Monica CA: Human Factors Society, 1991:62–6.
15. Shannon RH, Waag WL. Toward the development of a criterion for fleet effectiveness in the F-4 fighter community. Pensacola, FL: Naval Aeromedical Research Laboratory, 1972; NAMRL-1173.
16. Shannon RH, Waag WL. The prediction of pilot performance in the F-4 aircraft. Pensacola, FL: Naval Aeromedical Research Laboratory, 1973; NAMRL-1186.
17. Strawbridge D, Kahn N. Fighter pilot performance in Korea. Chicago, IL: Institute for Air Weapons Research, 1955; IAWR 55-10.
18. Tenney YJ, Adams JJ, Pew RW, Huggins AWF, Rogers WH. A principled approach to the measurement of situation awareness in commercial aviation. Langley, VA: National Aeronautics and Space Administration; 1992: NASA Contractor Report 4451.
19. Thomas GS, Houck MR, Bell HH. Training evaluation of air combat simulation. Brooks AFB, TX: Air Force Human Resources Laboratory, 1990; AFHRL-TR-90-30.
20. Torrance PE, Rush CH, Kohn HB, Doughty JM. Factors in fighter-interceptor pilot combat effectiveness. Lackland AFB, TX: Air Force Personnel and Training Research Center, 1957; AFPTRC-TN-57-23.
21. Waddell D. Situational awareness. USAF fighter weapons review, 1979; 27:3–5.
22. Youngling EW, Levine SH, Mocharnuk JB, Weston LM. Feasibility study to predict combat effectiveness for selected military roles: fighter pilot effectiveness. St. Louis, MO: McDonnell Douglas Astronautics Co., 1977; MDC E1634.